

Indraprastha Institute of Information Technology Delhi (IIITD)

Department of Computational Biotechnology

BIO213 – Introduction to Quantitative Biology

## ASSIGNMENT-1

**Name-PARAS DHIMAN || Roll no-2021482**

1.)

a.) To compute the global alignment using dynamic programming, we can use a matrix to store the scores of all possible alignments between the two sequences. The matrix is initialized with zeros, and the score of each cell  $(i,j)$  is calculated based on the scores of its adjacent cells as follows:

The score of a cell  $(i,j)$  is the maximum of the three possible values:

The score of the cell above  $(i-1,j)$  plus the gap penalty.

The score of the cell to the left  $(i,j-1)$  plus the gap penalty.

The score of the cell diagonally above and to the left  $(i-1,j-1)$  plus the match/mismatch score.

The optimal alignment can be traced back from the bottom right corner of the matrix to the top left corner by following the path of the maximum scores.

Using the given scoring scheme, we can compute the bidimensional array as follows

```
[0, -1, -2, -3, -4, -5, -6, -7, -8, -9]
[-1, 2, 1, 0, -1, -2, -3, -4, -5, -6]
[-2, 1, 0, -1, 2, 1, 0, -1, -2, -3]
[-3, 0, -1, -2, 1, 4, 3, 2, 1, 0]
[-4, -1, 2, 1, 0, 3, 2, 1, 4, 3]
[-5, -2, 1, 0, 3, 2, 5, 4, 3, 6]
[-6, -3, 0, 3, 2, 1, 4, 3, 2, 5]
[-7, -4, -1, 2, 1, 0, 3, 2, 5, 4]
[-8, -5, -2, 1, 0, -1, 2, 1, 0, -1]
[-9, -6, -3, 0, -1, 2, 1, 0, -1, 2]
```

In this matrix, each cell  $(i,j)$  contains the score of the best alignment up to that point.

b.) Yes there is possibility of more than one optimally aligning the given sequences:

Yes, there will be more than one optical alignment

This is because there can be multiple paths that result in the same maximum score.

c.) To find all optimal alignments, we can start at the bottom right corner of the matrix and trace back all the paths that lead to the top left corner, keeping track of the sequences as we go.

All the Optical alignments that we get are:

```
G--GCAGT-A-
GATGC-G-CAG
The score is: 4
-----
G--GCAG-TA-
GATGC-GC-AG
The score is: 4
-----
---G-GCAGTA
GATGCGCAG--
The score is: 4
-----
G----GCAGTA
GATGCGCAG--
The score is: 4
-----
G--G--CAGTA
GATGCGCAG--
The score is: 4
-----
G--GC--AGTA
GATGCGCAG--
The score is: 4
-----
Total Alignments are 6
```

2.)

Yes, changing the scoring scheme will modify the results obtained in Question 1. Using the new scoring scheme of [Match = +2, Mismatch = -1, Gap = -3], we can compute the bidimensional array as follows:

a.)

[0, -3, -6, -9, -12, -15, -18, -21, -24, -27]
[-3, 2, -1, -4, -7, -10, -13, -16, -19, -22]
[-6, -1, 1, -2, -2, -5, -8, -11, -14, -17]
[-9, -4, -2, 0, -3, 0, -3, -6, -9, -12]
[-12, -7, -2, -3, -1, -3, -1, -4, -4, -7]
[-15, -10, -5, -3, -1, -2, -1, -2, -5, -2]
[-18, -13, -8, -3, -4, -2, -3, -2, -3, -5]
[-21, -16, -11, -6, -4, -5, -3, -4, 0, -3]

b.) No, there is only one possibility

No, There is only one Optimal alignment possible

c.) To find the optimal alignments, we can start at the bottom right corner of the matrix and trace back all the paths that lead to the top left corner, keeping track of the sequences as we go.

```
G--GCAGTA-
GATGC-GCAG
The score is: -3
-----
Total Alignments are 1
```

Therefore, changing the scoring scheme results in a different optimal alignment with a lower score. This shows that the choice of scoring scheme can significantly impact the results of sequence alignment.

3.)

To perform local pairwise sequence alignment, we need to modify the dynamic programming algorithm by disallowing negative scores and identifying the optimal local alignment within the matrix. This is achieved through the use of the Smith-Waterman algorithm.

a.)

Using the Smith-Waterman algorithm we generate the following bidimensional array:

[0, 0, 0, 0, 0, 0, 0, 0, 0, 0]
[0, 2, 0, 0, 2, 0, 2, 0, 0, 2]
[0, 2, 1, 0, 2, 1, 2, 1, 0, 2]
[0, 0, 1, 0, 0, 4, 1, 4, 1, 0]
[0, 0, 2, 0, 0, 1, 3, 1, 6, 3]
[0, 2, 0, 1, 2, 0, 3, 2, 3, 8]
[0, 0, 1, 2, 0, 1, 0, 2, 1, 5]
[0, 0, 2, 0, 1, 0, 0, 0, 4, 2]

b.)

To find the optical alignment, we start from the block with the maximum score in the matrix and then trace back all the paths leading to the top row/column, i.e., those with an index of 0

All the Optical alignments that we get are:

```
GCAG
GCAG
The score is: 8
-----
Total Alignments are 1
```

4.)

To perform local pairwise sequence alignment instead of global pairwise sequence alignment, the following changes are required in the program:

Scoring system: The scoring system should be modified to not allow for negative scores. This is because in local alignment, we want to find regions of high similarity, even if they are surrounded by regions of dissimilarity. Negative scores allow for mismatches and gaps in the alignment.

```

highest = -1
idx1 = []
idx2 = []

for i in range (0,row + 1):
    for j in range(0,col + 1):
        if (lst[i][j] > highest):
            highest = lst[i][j]

for i in range (0,row + 1):
    for j in range(0,col + 1):
        if (lst[i][j] == highest):
            idx1.append(i)
            idx2.append(j)

```

Alignment initiation: In local alignment, the alignment is initiated by selecting a substring from each sequence, rather than aligning the entire sequences as in global alignment. The substring can be selected randomly, or by using an algorithm that identifies potential regions of similarity.

Dynamic programming: The dynamic programming algorithm used for local alignment is different from the one used for global alignment. The algorithm for local alignment is called the Smith-Waterman algorithm. This algorithm uses a similar matrix as the Needleman-Wunsch algorithm for global alignment, but with some modifications.

Backtracking: After the dynamic programming matrix is filled, the alignment can be obtained by backtracking through the matrix starting from the highest-scoring cell. In local alignment, the traceback starts at the cell with the highest score in the matrix and continues until a cell with a score of 0 is reached. This will give the best local alignment score and the corresponding alignment.

Overall, the main changes required for local pairwise sequence alignment involve modifications to the scoring system, alignment initiation, dynamic programming, and backtracking.